TECHNO INTERNATIONAL NEW TOWN

(Formerly known as Techno India College of Technology)

Block-DG, Action Area 1, New Town, Kolkata -700156, West Bengal, India

Department of Computer Science & Engineering

Eight Semester Project-III Report (PROJ-CS881)

Predicting diabetes with the help of Data Analytics and Machine Learning

Prepared by

Rajarshi Baral (Roll No:-18700122204)

Aoyan Mondal (Roll No:- 18700121024)

Sangeeta Barua (Roll No:-18700122196)

Arpita Saha (Roll No:-18700122187)

Under the Guidance of

Prof. Mr. Swarup Chakraboty

Batch:- 2021-2025 Semester :8 th (2025 –EVEN) Year : January 2025 – June 2025

Stream:- Computer Science & Engineering Year of Study: 4th

Affiliated to

MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY, WESTBENGAL

(FORMERLY KNOWN AS WEST BENGAL UNIVERSITY OF TECHNOLOGY)

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to **Mr. Swarup Chakraborty** of the Department of Computer Science & Engineering, whose role as project guide was invaluable for the project. We are extremely thankful for the keen interest he took in advising us, for the books and reference materials provided for the moral support extended to us.

Last but not the least we convey our gratitude to all the teachers for providing us the technical skill that will always remain as our asset and to all non-teaching staff for the cordial support they offered.

Place: Techno International New Town

Date: _ _ _ _ _ _ _ _ _ _ _

Rajarshi Baral
(Roll No: - 18700122204)

Aoyan Mondal
(Roll No: - 18700121024)

Sangeeta Barua
(Roll No: - 18700122196)

Arpita Saha
(Roll No: - 18700122187)

Department of Computer Science & Engineering,
Techno International New Town
Kolkata – 700 156
West Bengal, India.

Approval

This is to certify that the project report entitled ***"Predicting Diabetes Using Data Analytics and Machine Learning"*** prepared under my supervision by **Rajarshi Baral (18700122204), Aoyan Mondal (18700121024), Sangeeta Barua (18700122196), and Arpita Saha (18700122187)** , be accepted in partial fulfillment for the degree of Bachelor of Technology in Computer Science & Engineering which is affiliated to Maulana Abul Kalam Azad University of Technology, West Bengal (Formerly known as West Bengal University of Technology).

It is to be understood that by this approval, the undersigned does not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn thereof, but approves the report only for the purpose for which it has been submitted.

…………………………………………

Prof. Mr. Swarup Chakraborty

………………………………………….

Dr. Swagata Paul,
HOD, Computer Science & Engineering,
Techno International New Town

Abstract

This project focuses on predicting diabetes using data analytics and machine learning techniques implemented in Python. By analyzing health-related data such as glucose levels, BMI, age, and other medical parameters, the model identifies patterns to assess the likelihood of diabetes in individuals. Various Python packages like Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn were used for data preprocessing, visualization, and building predictive models. Several machine learning algorithms, including Logistic Regression, Decision Tree, and Random Forest, were applied and evaluated for accuracy. The project demonstrates the effectiveness of machine learning in enhancing early diagnosis and supporting data-driven healthcare decisions. The results highlight the potential of machine learning in improving diagnostic accuracy and supporting clinical decision-making. This project demonstrates how data-driven methods can contribute to the early detection of diabetes, ultimately aiding in better disease management and prevention strategies.

CONTENTS

1. INTRODUCTION 1
2. PROBLEM DEFINITION 2
3. BACKGROUND/SURVEY 3
4. PROPOSED METHODOLOGY 4-8
   4.1 Architecture /Flow Diagram 4
   4.2 State Diagram 5
   4.3 Machine Learning Algorithm 5-9
5. REQUIREMENT SPECIFICATION 10-14
6. OUTPUT /RESULT ANALYSIS 15-17
7. FUTURE SCOPE OF PROJECT 18
8. CONCLUSION 19
9. BIBLIOGRAPHY 20

1. Introduction

Diabetes is one of the most prevalent chronic diseases affecting millions of people worldwide. Early detection and proper management are crucial to reducing the risk of severe complications associated with the disease. In recent years, advancements in data science and machine learning have opened new pathways for improving the accuracy and efficiency of medical diagnoses, including the prediction of diabetes.

This project focuses on the use of data analytics and machine learning techniques, implemented through Python programming language and its powerful libraries such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn, to develop a predictive model for diabetes. By analyzing historical medical data and identifying key risk factors such as glucose levels, BMI, age, and blood pressure, the system aims to classify individuals as diabetic or non-diabetic.

The objective of this project is not only to demonstrate the predictive capabilities of machine learning algorithms such as Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines (SVM) but also to emphasize the importance of data preprocessing, visualization, and evaluation metrics like accuracy, precision, recall, and F1-score.

Through this approach, we aim to contribute toward more effective, data-driven healthcare solutions that can assist medical professionals in making informed decisions, thereby improving patient outcomes and raising awareness about the risk factors associated with diabetes.

2. Problem Definition

Diabetes is a chronic health condition that poses a significant challenge to global healthcare systems. Early detection and proper management are crucial to preventing complications and improving patient outcomes. However, traditional diagnostic methods can be time-consuming and may not always yield accurate results, especially in the early stages of the disease.

With the increasing availability of medical data, there is a growing opportunity to apply data analytics and machine learning techniques to enhance the prediction and diagnosis of diabetes. The primary objective of this project is to develop a predictive model that can accurately identify the likelihood of an individual being diabetic based on various medical and lifestyle parameters.

This project utilizes Python as the core programming language along with widely used libraries such as Pandas, NumPy, Matplotlib, Seaborn for data preprocessing and visualization, and Scikit-learn, XGBoost, and TensorFlow/Keras for building and evaluating machine learning models.

The model aims to:

- Analyze and process real-world medical datasets (e.g., the Pima Indians Diabetes Dataset).
- Identify the most significant features contributing to diabetes prediction.
- Apply and compare different classification algorithms to determine the most accurate approach.
- Provide a reliable and efficient tool that can assist healthcare professionals in early diabetes prediction.

The outcome of this project is expected to contribute to data-driven healthcare and offer an intelligent support system for early diagnosis, thereby aiding timely medical intervention and improving patient care.

3. BACKGROUND/SURVEY

Diabetes mellitus is a chronic and potentially life-threatening disease that affects millions of individuals worldwide. The early detection and management of diabetes are crucial in preventing complications and improving quality of life. Traditional diagnostic methods often involve invasive procedures and may not always detect the condition in its early stages. With the increasing availability of medical data and advancements in computational techniques, data-driven approaches have gained prominence in healthcare.

Machine learning (ML), a subset of artificial intelligence, offers powerful tools for analyzing large datasets and making accurate predictions. When integrated with data analytics, ML models can identify complex patterns and correlations in health data that may not be evident through conventional statistical techniques. Python, being a versatile and open-source programming language, has become the preferred platform for implementing machine learning algorithms due to its extensive libraries and frameworks like Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn.

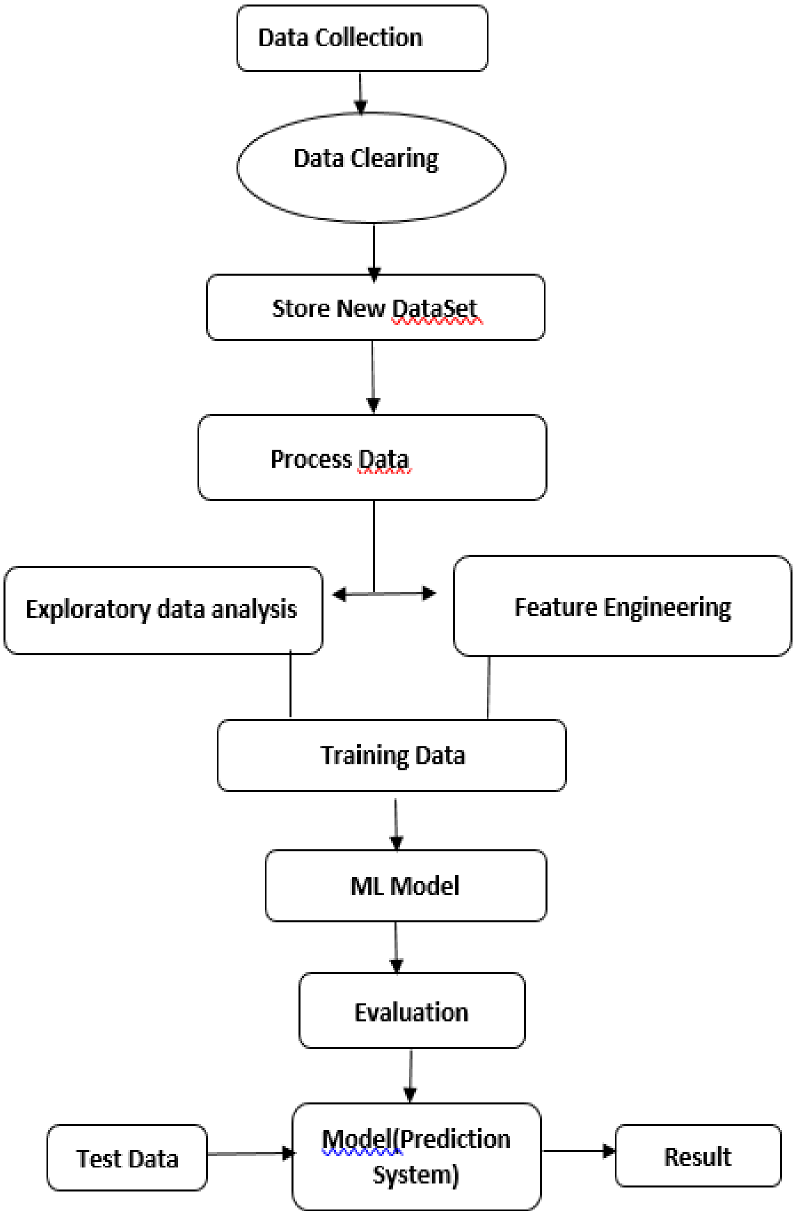
Numerous studies and projects have demonstrated the effectiveness of ML techniques—such as logistic regression, decision trees, support vector machines (SVM), and random forests—in predicting diabetes. These models are typically trained on datasets such as the Pima Indians Diabetes Dataset, which includes medical parameters like glucose level, BMI, age, blood pressure, and insulin levels.

The use of Python in this domain allows for seamless data preprocessing, feature selection, model training, and evaluation. Data analytics plays a critical role in understanding the data distribution, handling missing values, and visualizing patterns, which further enhances model performance.

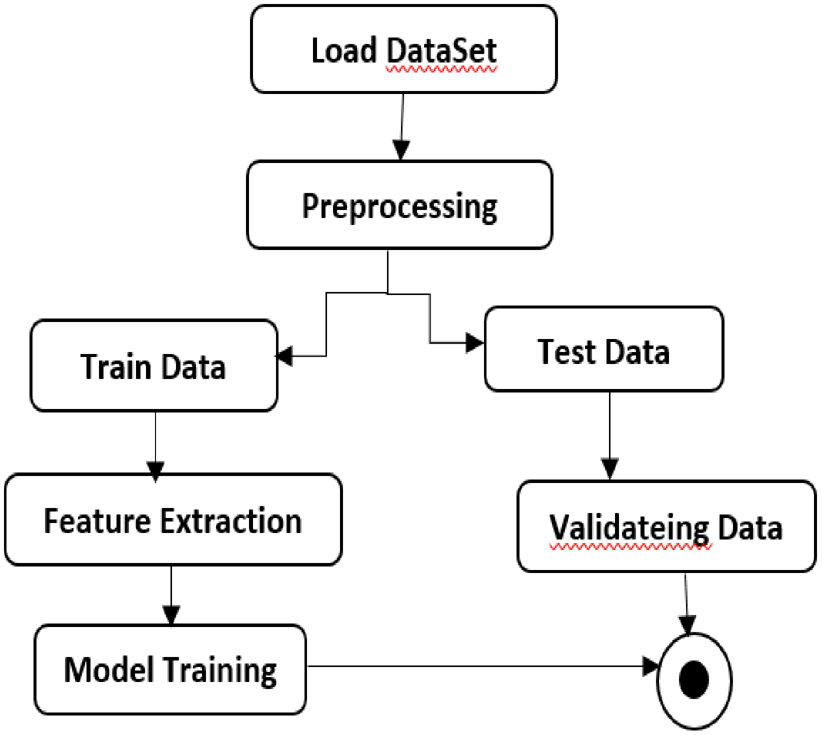
This project aims to leverage Python-based machine learning algorithms and data analytics techniques to develop an effective diabetes prediction model. The objective is to enable early diagnosis by accurately classifying whether a patient is likely to develop diabetes based on their medical attributes, thus aiding in timely medical intervention.

4. Proposed Methodology

a. Architecture/Flow Diagram:

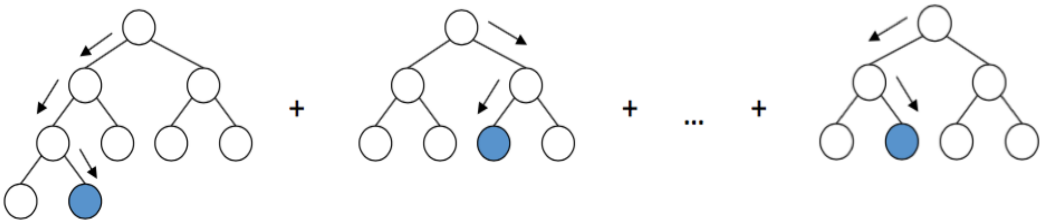


b. **State Diagram:**

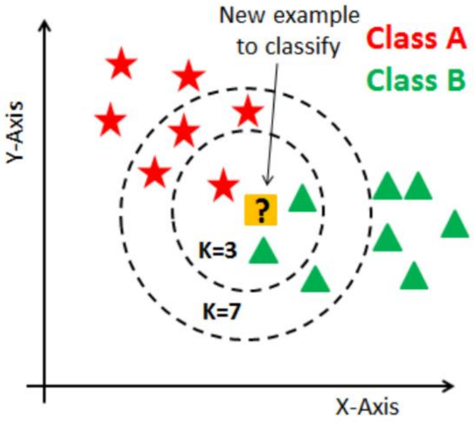


- **Machine Learning Algorithm:** We have also use various algorithms in our project to train the model and they are following: 1) Gradient Boosting Machine
  2) K-Nearest Neighbors
  3) Logistic Regression
  4) Random Forest
  5) Support Vector Machine

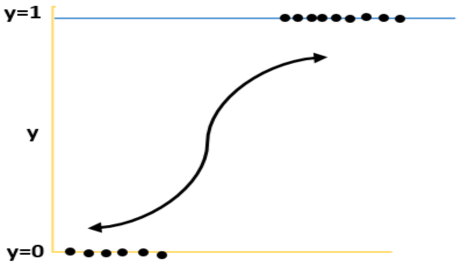
i. **Gradient Boosting Machine (GBM): Gradient Boosting** is a powerful supervised machine learning algorithm used for both classification and regression problems. It works on the principle of ensemble learning, where multiple weak learners (typically decision trees) are combined to form a strong predictor. Unlike Random Forest, which builds trees in parallel, Gradient Boosting builds them sequentially — each new tree corrects the errors made by the previous ones. The process starts with an initial prediction (like the average value), and then the algorithm computes the residual errors — the difference between the predicted and actual values. A new tree is trained to predict these residuals. This process is repeated for a set number of iterations, and the final model is a weighted sum of all trees.Gradient Boosting uses gradient descent to minimize a loss function (such as log loss for classification), hence the name. It is highly flexible and can be tuned with parameters like learning rate, tree depth, and number of trees.In diabetes prediction, Gradient Boosting is very effective because it captures complex patterns in patient data (like glucose level, BMI, blood pressure) and gives accurate results. However, it can be computationally intensive and prone to overfitting if not properly tuned.



ii. **K-NEAREST NEIGHBORS:** The **K-Nearest Neighbors (K-NN)** algorithm is a popular and simple supervised learning method used for both classification and regression tasks. It is based on the idea that data points with similar characteristics are likely to belong to the same class or have similar output values. The algorithm does not build an explicit model but instead stores all the training data. When a new data point needs to be predicted, K-NN calculates the distance between this point and all other points in the training dataset. The most common distance metric used is **Euclidean distance**, though others like Manhattan or Minkowski can also be used.After calculating the distances, the algorithm selects the **K closest neighbors** (where K is a user-defined constant) and uses them to determine the result. In classification, the new point is assigned the class that is most common among its K nearest neighbors. In regression, the predicted value is the average of the values of these neighbors. K-NN is intuitive and easy to implement but can be computationally expensive for large datasets, especially during the prediction phase. It is sensitive to the choice of K, irrelevant features, and the scale of data. Therefore, **feature normalization** and careful **K selection** are crucial for good performance.



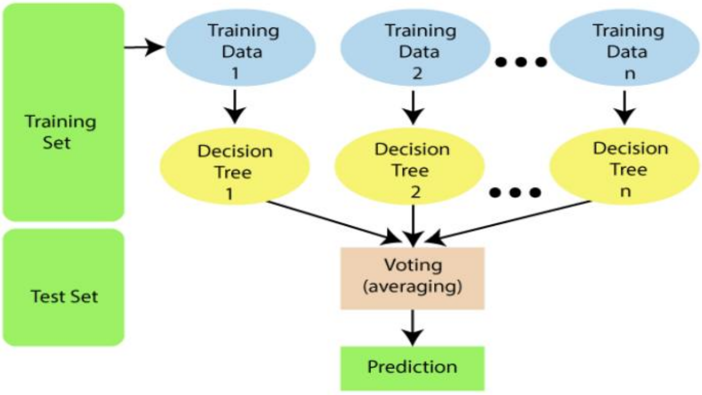
iii. **Logistic Regression: Logistic Regression** is a widely used supervised machine learning algorithm primarily used for binary classification problems, such as predicting whether a patient has diabetes (Yes/No). Unlike linear regression, which predicts continuous values, logistic regression predicts the probability that a given input belongs to a certain class. The core of logistic regression is the logistic (sigmoid) function, which transforms any real-valued number into a value between 0 and 1. This value represents the probability of the target class. If the probability is greater than 0.5, the output is usually classified as class 1 (e.g.,

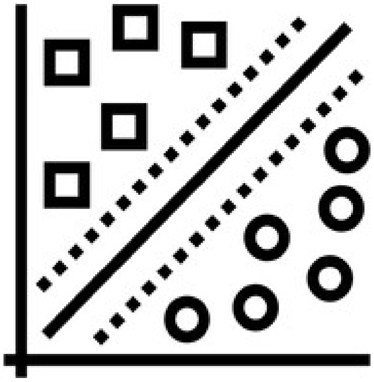


diabetic); otherwise, it is class 0 (non-diabetic).Logistic regression uses input features such as glucose level, BMI, age, and blood pressure, and assigns weights to these features to estimate the likelihood of the outcome. During training, the algorithm adjusts these weights to minimize the log loss, a measure of prediction error. In healthcare, especially in diabetes prediction, logistic regression is valuable due to its predictive power, interpretability, and clinical reliability.

iv. **Random Forest: Random Forest** is a powerful and widely used ensemble learning algorithm in machine learning, primarily used for classification and regression tasks. It builds multiple decision trees during training and merges their outputs to improve prediction accuracy and control overfitting. In a Random Forest, each tree is trained on a random subset of the data (with replacement, known as bagging) and uses a random subset of features at each decision split. This randomness makes the forest diverse and more robust compared to a single decision tree. During prediction, each tree in the forest gives a result (e.g., diabetic or non-diabetic), and the final output is decided by majority voting (for classification) or averaging (for regression). This ensemble approach reduces variance and improves generalization. Random Forest is especially useful in diabetes prediction, as it can handle large datasets with many features and detect complex, non-linear relationships between patient health indicators such as glucose level, BMI, insulin, and blood pressure. It is resistant to overfitting, handles missing data well, and provides feature importance scores, helping identify which medical attributes are most predictive. Due to its high accuracy, robustness, and interpretability, Random Forest is a popular choice in healthcare analytics and predictive modeling.

v. **Support Vector Machine: Support Vector Machine (SVM)** is a powerful supervised machine learning algorithm used for both classification and regression tasks, but it is especially effective for binary classification problems like diabetes prediction. SVM works by finding the optimal hyperplane that best separates the data points of different classes (e.g., diabetic vs. non-diabetic) in a high-dimensional space. The goal is to maximize the margin between the two classes—the distance between the hyperplane and the nearest data points from each class, known as support vectors. SVM can handle both linearly separable and non-linearly separable data. For complex datasets, it uses kernel functions (like polynomial, RBF, or sigmoid) to transform the input space into a higher dimension where a linear separator can be applied. In the context of diabetes prediction, SVM uses patient features such as glucose level, BMI, insulin, and age to classify whether a patient is likely to have diabetes. It is particularly useful when the data has clear margins of separation and is not too large. SVM offers high accuracy, especially in medical datasets, and works well even with limited data. However, it can be computationally intensive and less interpretable compared to simpler models like logistic regression.





5. Requirement Specifications

5.1 SYSTEM CONFIGURATION:

- HARDWARE REQUIREMENTS: Processer : Any Updated Processer
- RAM : Min 4GB Hard Disk : Min 100GB
- SOFTWARE REQUIREMENTS: Operating System
- WINDOWS FAMILY TECHNOLOGY : Python3.6
- IDE: Jupiter notebook

5.2 HARDWARE PLATFORM USED: The hardware requirement may serve as the basis for a contract for the implementation of the system and should therefore be complete and consistent in specification. The hardware used for the system is mentioned below.

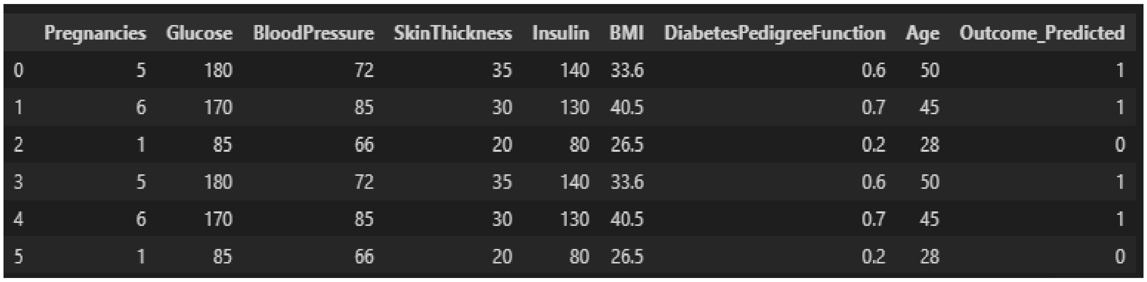
- PROCESSOR: Intel CORE i3 or above
- RAM: minimum 4.00GB
- HARD DISK: minimum 100GB It should be noted that better the hardware facilities available, higher wouldbe response time of the system.

5.3 LIBRARIES AND SOFTWARE PLATFORM USED: The software requirement document is the specification of the system. The software requirement provides a basis for creating the software requirements specification.

- OPERATING SYSTEM: Windows 11
- SYSTEM TYPE: 64-bit, intel CORE i5
- SOFTWARE: Jupyter Notebook, VS Code
- TECHNOLOGIES: Python
- LIBRARIES: pandas, numpy, scikit-learn, matplotlib, seaborn, streamlit, joblib etc.

5.4 Sample Code: -

5.4.1 DATASET DETAILS: Dataset has been obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. Diabetes dataset has 9 attributes in total. All the person in records are females and the number of pregnancies they have had has been recorded as the first attribute of the dataset. The dataset contains 768 records of female patients.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome_Predicted
0	5	180	72	35	140	33.6	0.6	50	1
1	6	170	85	30	130	40.5	0.7	45	1
2	1	85	66	20	80	26.5	0.2	28	0
3	5	180	72	35	140	33.6	0.6	50	1
4	6	170	85	30	130	40.5	0.7	45	1
5	1	85	66	20	80	26.5	0.2	28	0

Fig 5.1 Dataset Attributes

➢ Input dataset attributes

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- Body Mass Index(BMI)
- Diabetes Pedigree Function
- Age
- Outcome Predicted

5.4.2 **DISTRIBUTION OF DIABETIC PATIENT:** In our attempt to develop a diabetes prediction model, we encountered a slightly imbalanced dataset. Out of the total 768 samples, around 500 were Designated as 0, denoting the nonexistence of diabetes., while 268 were designated as 1, denoting the existence of diabetes.

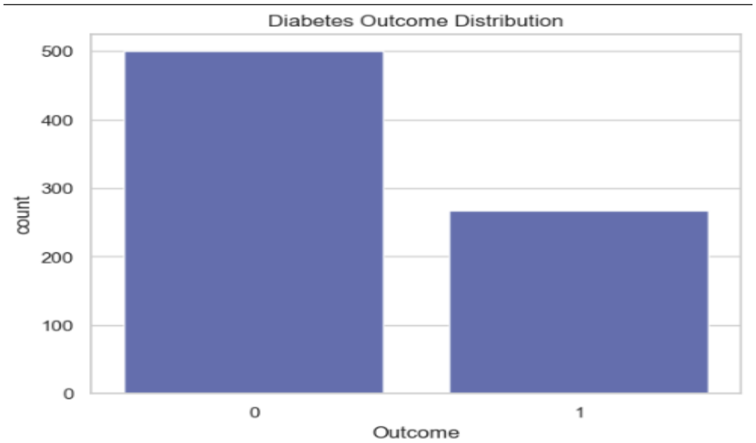
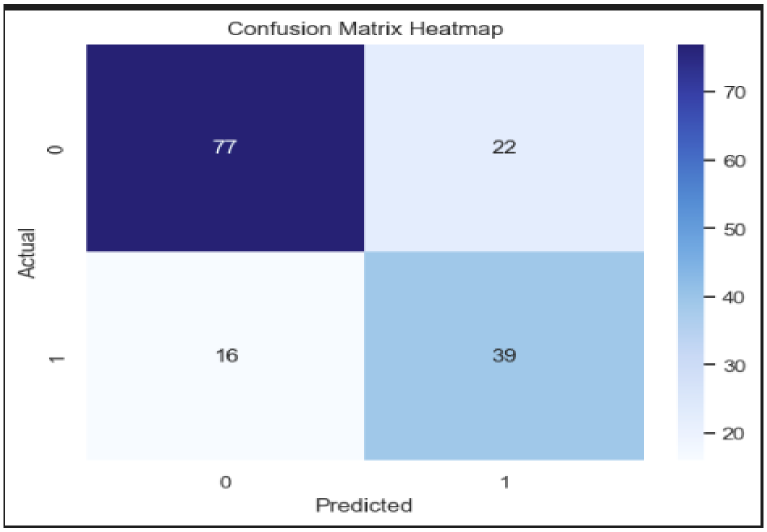


Fig 5.2 The proportion of patients with diabetes compared to those without diabetes

5.4.3 PERFORMANCE ANALYSIS: In this project, various machine learning algorithms like SVM, Decision Tree, Random Forest, Logistic a used to predict diabetes. Diabetes Prediction UCI dataset, has a total of 9 attributes, out of those only 9 attributes are considered for the prediction of Diabetes Prediction. Various attributes of the patient like Pregnancies, Glucose , Blood Pressure, Skin Thickness , Insulin etc. are considered for this project. The accuracy for individual algorithms has to measure and whichever algorithm is giving the best accuracy, that is considered for the diabetes prediction. For evaluating the experiment, various evaluation metrics like accuracy, confusion matrix, precision, recall, and f1-score are considered. Accuracy- Accuracy is the ratio of the number of correct predictions to the total number of inputs in the dataset. It is expressed as: Accuracy = (TP + TN) /(TP+FP+FN+TN),

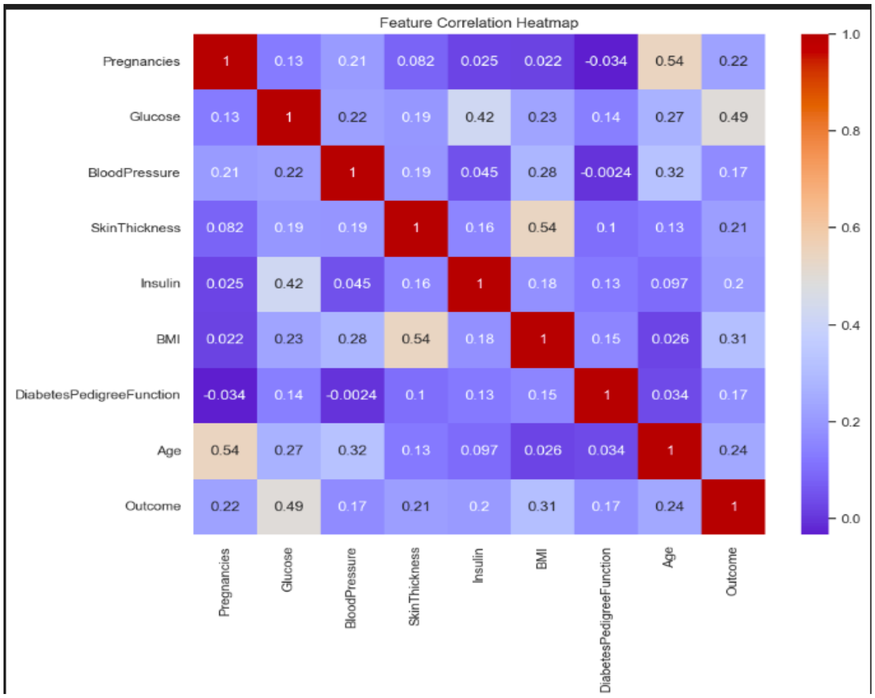
Confusion Matrix: It gives us a matrix as output and gives the total performance of the system.



Actual \ Predicted	0	1
0	77	22
1	16	39

Fig 5.3 Confusion Matrix Heatmap

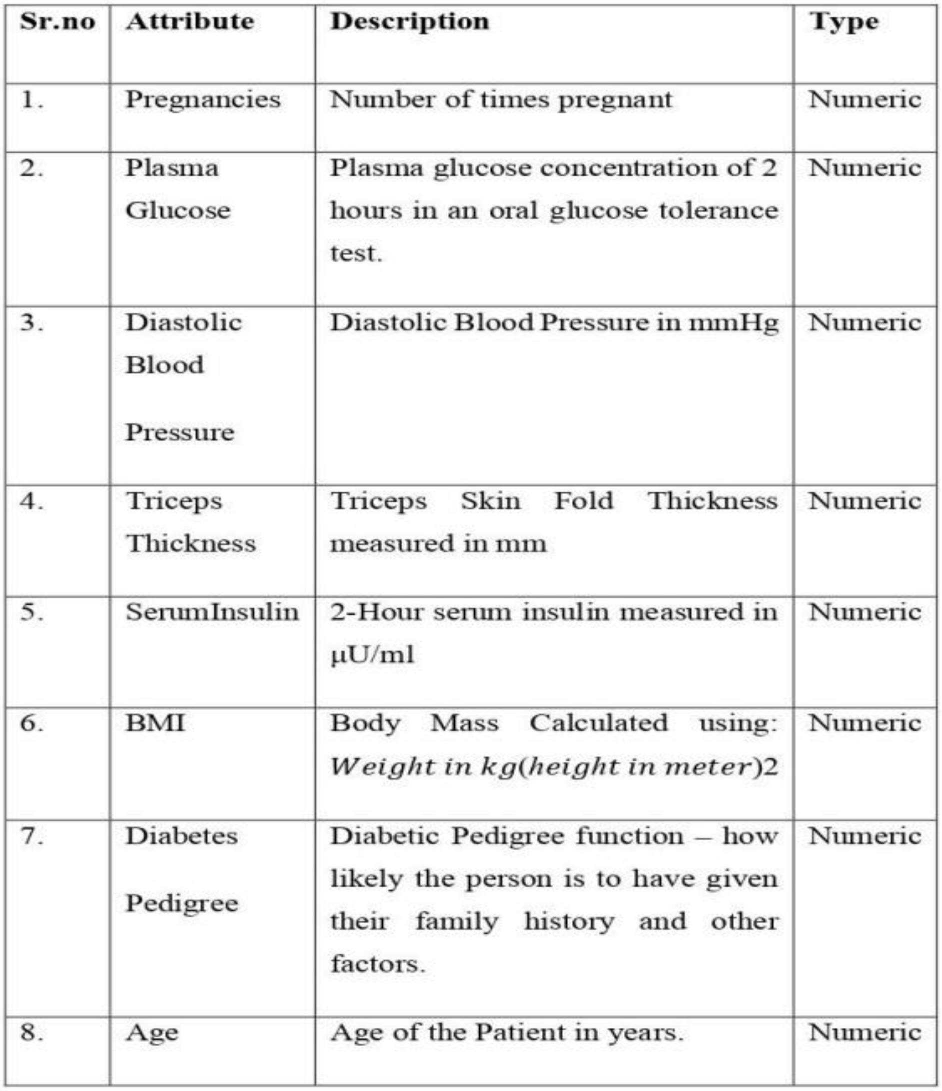
Correlation Matrix: The correlation matrix in machine learning is used for feature selection. It represents dependency between various attributes.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1	0.13	0.21	0.082	0.025	0.022	-0.034	0.54	0.22
Glucose	0.13	1	0.22	0.19	0.42	0.23	0.14	0.27	0.49
BloodPressure	0.21	0.22	1	0.19	0.045	0.28	-0.0024	0.32	0.17
SkinThickness	0.082	0.19	0.19	1	0.16	0.54	0.1	0.13	0.21
Insulin	0.025	0.42	0.045	0.16	1	0.18	0.13	0.097	0.2
BMI	0.022	0.23	0.28	0.54	0.18	1	0.15	0.026	0.31
DiabetesPedigreeFunction	-0.034	0.14	-0.0024	0.1	0.13	0.15	1	0.034	0.17
Age	0.54	0.27	0.32	0.13	0.097	0.026	0.034	1	0.24
Outcome	0.22	0.49	0.17	0.21	0.2	0.31	0.17	0.24	1

Fig 5.4 Correlation Matrix

Precision: It is the ratio of correct positive results to the total number of positive results predicted by the system. It is expressed as: Recall-It is the ratio of correct positive results to the total number of positive results predicted by the system. It is expressed as: F1 Score-It is the harmonic meaning of Precision and Recall. It measures the test

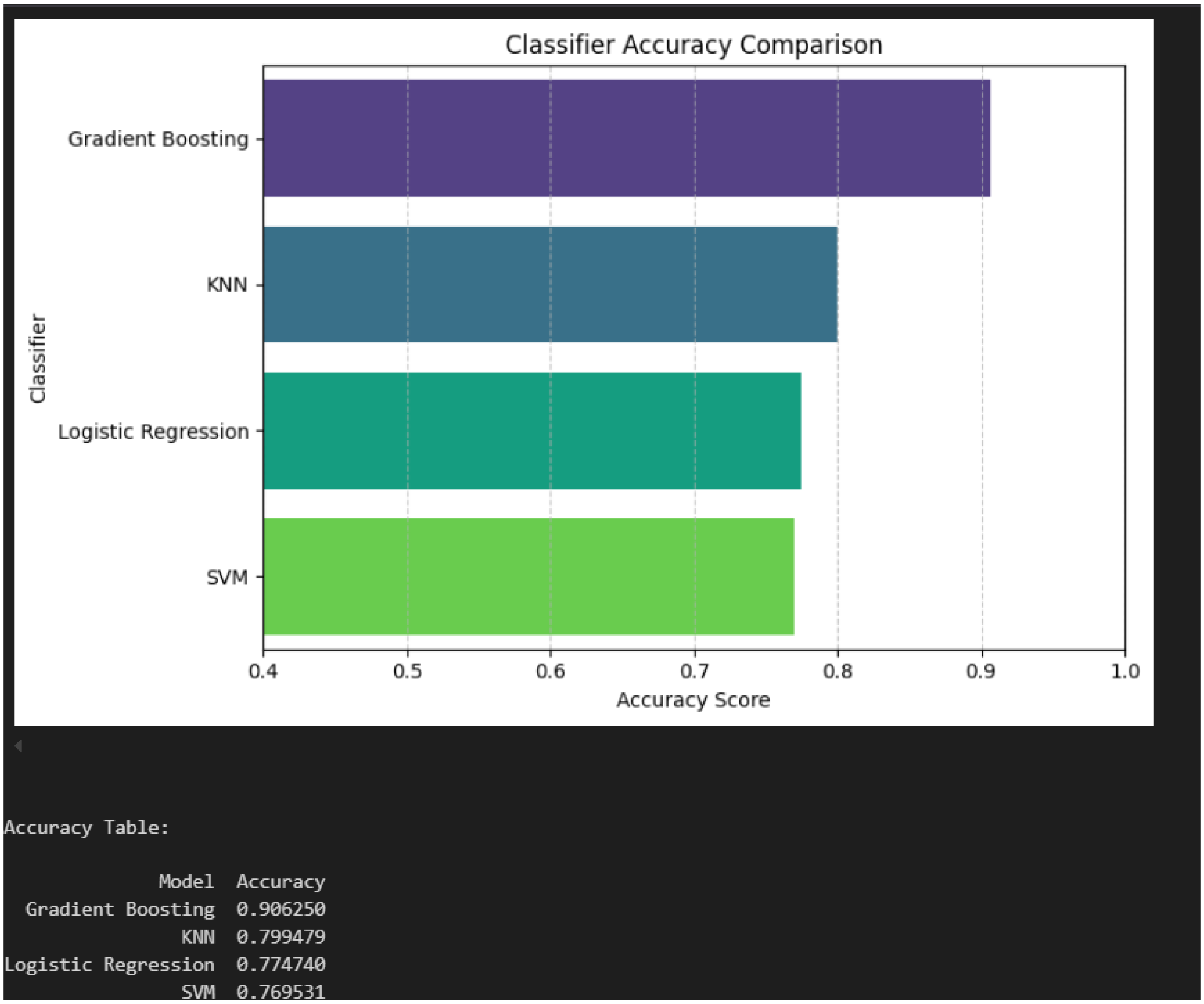


Sr.no	**Attribute**	**Description**	**Type**
1.	Pregnancies	Number of times pregnant	Numeric
2.	Plasma Glucose	Plasma glucose concentration of 2 hours in an oral glucose tolerance test.	Numeric
3.	Diastolic Blood Pressure	Diastolic Blood Pressure in mmHg	Numeric
4.	Triceps Thickness	Triceps Skin Fold Thickness measured in mm	Numeric
5.	SerumInsulin	2-Hour serum insulin measured in µU/ml	Numeric
6.	BMI	Body Mass Calculated using: $Weight\ in\ kg(height\ in\ meter)2$	Numeric
7.	Diabetes Pedigree	Diabetic Pedigree function – how likely the person is to have given their family history and other factors.	Numeric
8.	Age	Age of the Patient in years.	Numeric

Fig 5.4 Attribute Detrails

6. OUTPUT /RESULT ANALYSIS

The table below displays the performance values of various classification algorithms, calculated using different measures. Based on the table, it is observed that Logistic regression exhibits the highest accuracy. Therefore, the Logistic regression machine learning classifier is capable of predicting the likelihood of diabetes with greater precision than other classifiers.



```
Accuracy Table:

              Model  Accuracy
  Gradient Boosting  0.906250
                KNN  0.799479
Logistic Regression  0.774740
                SVM  0.769531
```

Fig 6.1 Classifier Accuracy Comparis

- Accurancy Table:

Model	Accuracy
Gradient Boosting	0.906250
KNN	0.799479
Logistic Regression	0.77470
SVM	0.769531

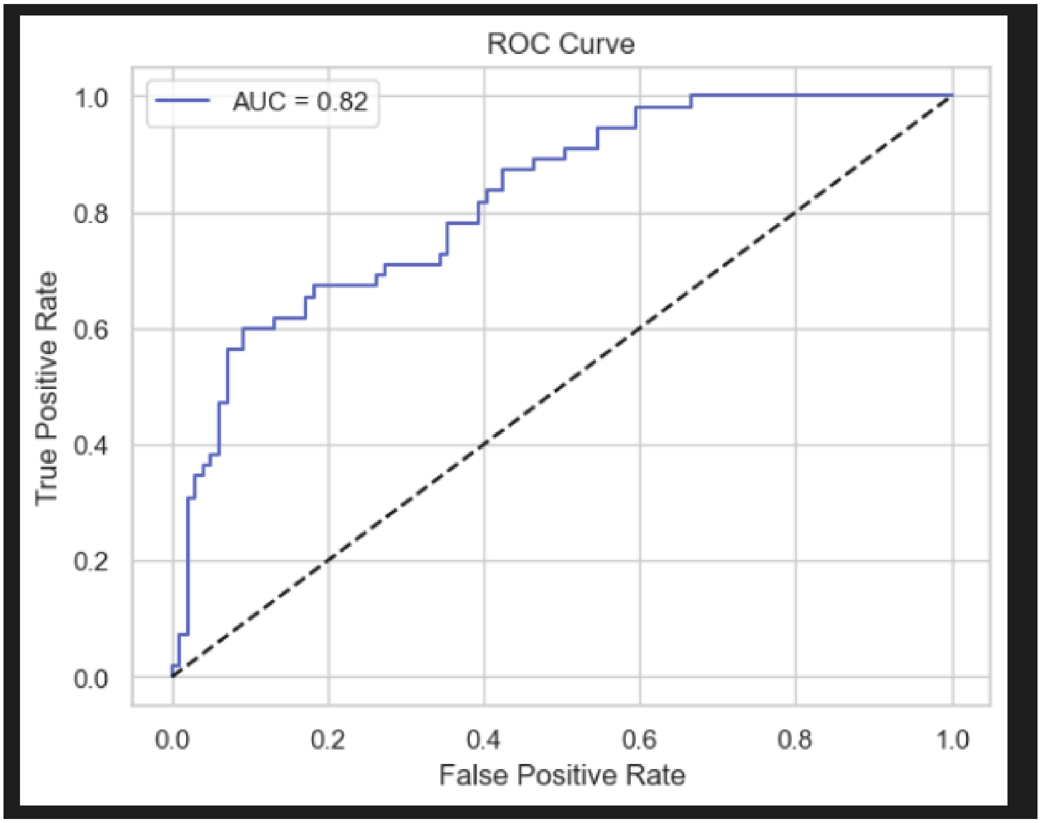


Fig 6.2 ROC Curve

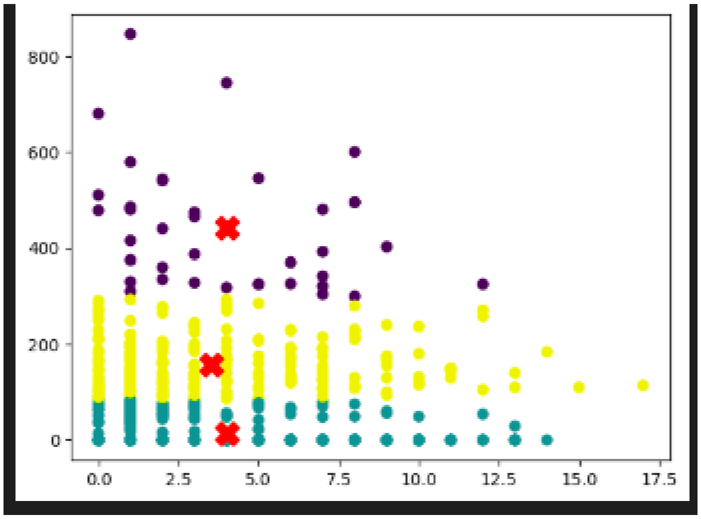


Fig 6.3K-means Mode

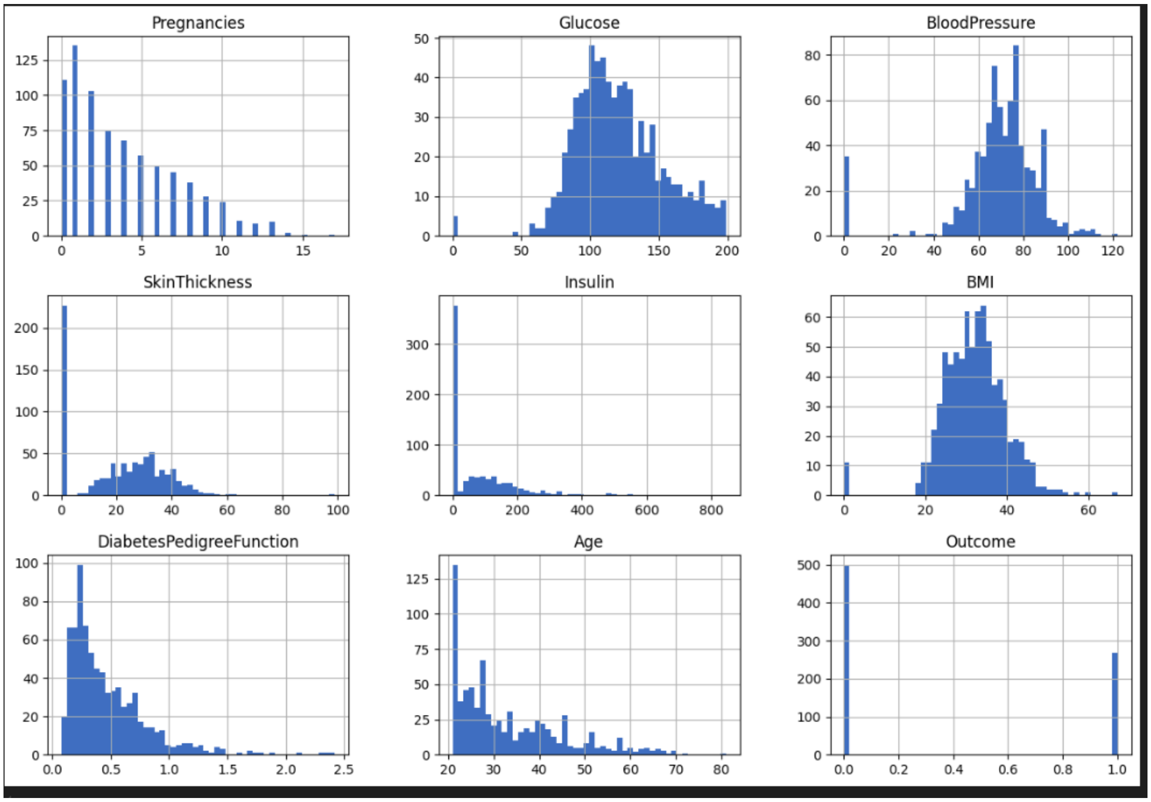


Fig 6.4 Pair plotting of data frame

7. Future Scope of the Project

In the future, we plan to create a diabetes dataset in collaboration with hospitals or medical institutions to achieve more accurate results. We also aim to incorporate additional machine learning and deep learning models to further improve prediction accuracy.

For future work, it is essential to collect real-time and up-to-date patient data from hospitals for the continuous training and enhancement of our current model. The dataset should be large enough to support reliable training and forecasting.

Moreover, advanced data mining techniques and visualizations should be explored to gain deeper insights. It is important to establish a framework of guidelines and quality standards to ensure ethical use of data mining, preventing misuse of sensitive health information.

This approach will help in slowing down the rise in blood glucose levels and ultimately reduce the risks associated with inaccurate predictions. Responsible data mining practices will play a crucial role in enhancing the effectiveness and reliability of diabetes prediction systems.

8. Conclusion

The primary aim of this project was to design and implement a **Diabetes Prediction System** using various **machine learning methods** and to analyze the performance of those methods. This goal has been successfully achieved.

The project focused on assessing the effectiveness of **Logistic Regression** compared to other linear classifiers such as **Support Vector Machine (SVM)**, **K-Nearest Neighbors (KNN)**, **Random Forest (RF)**, **Decision Tree**, and **Gradient Boosting**. The results of the comparison revealed that Logistic Regression outperformed the other classifiers, achieving the highest **classification accuracy of 83%**.

The proposed approach utilized **ensemble learning** and multiple classification techniques, which contributed to achieving high levels of prediction accuracy. These experimental results demonstrate the potential of machine learning in assisting healthcare professionals by enabling **early detection of diabetes** and supporting informed clinical decisions.

By applying these machine learning models, particularly Logistic Regression, the system can help in the **early diagnosis of diabetes**, which can lead to timely treatment and potentially **save human lives**. This project highlights the importance of integrating AI-driven models into healthcare systems for better patient outcomes.

9. Bibliography

- Arwatki Chen Lyngdoh, Nurul Amin Choudhury, Soumen Moulik, "Diabetes Disease Prediction Using Machine Learning Algorithms", IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES) 2020, DOI: 10.1109/IECBES48179.2021.9398759 .
- K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.
- M Sabibullah, V Shanmugasundaram and Priya K Raja, ”Diabetes Patient's Risk through Soft Computing Model", International Journal of Emerging Trends Technology in Computer Science, vol. 2, no. 6, 2013.
- A.K., Dewangan, and P., Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," International Journal of Engineering and Applied Sciences, vol. 2, 2015.
- **https://youtu.be/dMn2QFTyXUQ?si=Vh-kKZ7SurRx443h**
- **https://youtu.be/bno03RUhMIY?si=EI3IrpsEljnpBRfY**